

*DOI:*
10.1080/08913811.2016.1237704

*Document Version*
Peer reviewed version

# Not a New Gold Standard:

# Even Big Data Cannot Predict the Future

**Kai Jäger**
Department of Political Science, University of Mannheim
Mannheim Centre for European Social Research (MZES

**ABSTRACT**

Many scholars believe that the proliferation of large-scale datasets will spur scientific advancement and help us to predict the future using sophisticated statistical techniques. Indeed, a team of researchers achieved astonishing success using the world's largest event dataset, produced by the ICEWS project, to predict complex social outcomes such as civil wars and irregular government turnovers. However, the secret of their success lay in transforming epistemically difficult questions into easy ones. Forecasting the onset of civil wars becomes an easy task if one relies on explanatory variables that measure how often newspapers report on tensions, fights, or killings shortly before. But news reports on prewar conflicts are just variations of the variable that researchers want to predict; the finding that more conflicts are likely to occur when journalists report about conflicts carries little scientific value. A similar success rate in "predicting" interstate wars can also be achieved by a simple Google News search for country names and conflict-related news shortly before a conflict is coded as a war. Big data can help researchers to make predictions in simple situations, but there is no evidence that predictions will also succeed in uncertain environments with complex outcomes—such as those characteristic of politics.

**Keywords:** big data, event data, ICEWS, ideational factors, machine coding, Nassim Taleb, peace and conflict studies, Thai conflict, unpredictability.

*Almost all wars are preceded by episodes of tension; almost all episodes of tension are succeeded by peace.*

—Nassim Nicholas Taleb[1]

## 1. Introduction

The question of whether reliable predictions are possible in political science has been repeatedly discussed in *Critical Review*, particularly in two symposia "The Age of Uncertainty" (vol. 21, no. 4, 2009) and "Political Interaction" (vol. 24, no. 3, 2012). In his introductory article to the latter symposium, Jeffrey Friedman (2012, 309) wrote that "it does not take a leap of the imagination […] to 'predict' that political scientists might jump on the Nate Silver bandwagon and use statistics to try to predict things other than elections."

Especially in peace and conflict studies, attempts to predict political violence have become popular (e. g., Ulfelder 2012; Goldsmith et al. 2013; Hegre et al. 2013; Brandt et al. 2014). Friedman cited Michael Ward and Nils Metternich's *Foreign Policy* op-ed entitled "Predicting the Future Is Easier Than It Looks" (Ward and Metternich 2012) as prime example for the claim that statistical techniques will increasingly be able to predict events in world politics using disaggregated big-data collections. As the prediction project by Ward and colleagues is one of the largest and most influential, this article critically analyzes their apparent prediction success. After introducing the big-data project, I briefly discuss why it was unable to correctly represent conflicts in Thailand. Subsequently, I explain why it was still able to forecast political conflicts: By

[1] Facebook Entry, Nassim Nicholas Taleb, 2 February, 2015. https://www.facebook.com/permalink.php?story_fbid=10152794640733375&id=13012333374

transforming a difficult epistemic task into an easy one—specifically, by including variations of the dependent variable as explanatory variables.

## 2. Thailand as a Case Study in the Failure of Big-Data Prediction

Ward is part of Lockheed Martin's International Crisis Early Warning System (ICEWS), which is funded by U. S. government agencies. This large-scale project is designed to predict various types of conflicts in order to inform U. S. policy analysts in advance about new international hot spots. ICEWS utilizes daily news articles from the news aggregator Factiva and the U.S. Government's Open Source Center to construct a raw dataset of daily events for up to 167 countries (Ward et al. 2013a). The coding of the news reports is implemented by an automatic system, such as the Conflict and Mediation Event Observation (CAMEO) project, which uses predefined verbal and group categories to turn news reports into classified types of events (Schrodt and Van Brackle 2013). For the period January 2001 to April 2013, ICEWS consists of about 30 million coded news reports, resulting in a sample of about 16 million events, or an average of approximately 700 events for every country per month (Ward et al. 2013a, 4). It is probably the largest event dataset that is currently available.[2]

Based on the ICEWS dataset, sophisticated statistical techniques, and a game-theoretic approach, Ward and colleagues claim in several academic publications that they have revealed a gold standard for advancing political science by predicting the future. Ward (2016, 86) contends that their models "produce accurate forecasts, and as expected, the ensemble forecasts are

[2] The ICEWS dataset was recently made available for the public (Boschee et al. 2015).

generally more accurate, in terms of having both fewer false negatives and positives, than any individual model."

In a pilot study in the *American Journal of Political Science*, Ward, Metternich and other scholars use the ICEWS data and the CAMEO automatic coding scheme to measure conflicts and to construct network variables of anti-government groups in Thailand (Metternich et al. 2013). The authors claim to show that a fragmented anti-government network structure leads to more conflicts in Thailand. They also maintain that their network configuration performs better than a structural model in out-of-sample predictions of conflict frequency.

Elsewhere, I have argued that these conclusions are unfounded (Jäger 2016). For example, one of the worst episodes of violence in Thai history occurred during the period of the authors' out-of-sample prediction, when security forces cracked down on "red-shirt" protesters (supporters of former Prime Minister Thaksin) who had occupied central Bangkok for over two months between March and May 2010.[3] The ICEWS conflict variable wrongly portrayed this period as one of low and declining incidents of conflict. This is one of many ways in which big data proved utterly inadequate to the prediction of conflict in Thailand, contrary to the authors' claims. At bottom, these inadequacies stem from the inability of computers to understand the ideational context in which the actions they can correlate occur.

One of many glaring examples of the conflict between understanding political actors' ideas and the use of big-data statistical techniques in the Metternich et al. (2013) study concerns the coding of Thai political actors' groups in the ICEWS database. Coding is necessary in all statistical work of this kind in order to homogenize the qualitatively unique actions of a variety of people, so they can be counted up and compared to each other quantitatively. In checking the accuracy of the

[3] Official estimates count 85 deaths and 1,813 injured, while unofficial sources claim an even higher toll (Nidhi 2012, 14).

ICEWS coding, I used the following simple rules for data entries in the "anti-government" category: A group cannot be part of the anti-government network if it is actually in the government, if it has not been founded yet or has ceased to exist, or if it does not exist in Thai politics. Coded cooperative events between groups are also considered to be incorrect if they involve two groups that did not cooperate in the real world—for example, because they were enemies in Thai politics. Based on these simple criteria, my examination revealed that over half of the data entries in the ICEWS anti-government network are incorrect, and that the rate for incorrect cooperative events between groups approximates 80 percent. As a consequence, ICEWS's placement of groups in a two-dimensional political space—which is crucial for calculating the main network variable in Metternich et al.'s model—does not resemble the cleavages of real-world Thai politics. ICEWS fails to place allied Muslim insurgent groups together and inaccurately puts red (pro-Thaksin) and yellow (anti-Thaksin) groups closer to each other than to other parties. Indeed, the largest polarization ICEWS detects is between the pro-Thaksin Pheu Thai Party and its own social movement, the red shirts.

Another problem is that given the nature of their illegal activities, networks of insurgent groups have to try to avoid being detected by security forces or journalists. News reports are not a reliable source of information in this context, as insurgents would immediately abandon exposed network ties. By contrast, close relationships among political parties are too constant and well known over time to be considered newsworthy by journalists. For an accurate network analysis, however, news reports are required to regularly report the alliances of political groups. The systematic tendency of journalists to report only newsworthy information also affects the depiction of the relationships between antagonistic groups. A given conflictual relationship might be

constant over time, but journalists will cover it only when the conflict becomes "hot" and is in the spotlight.

As a consequence, any detected association between network activity and conflicts in the ICEWS dataset might be spurious, as it could simply reflect journalistic practices. The larger problem here is that big data has to originate in the actions of human interpreters of reality. In this case the interpreters are journalists, and—like all human interpreters—the actions they take (reporting what they consider to be newsworthy) produce patterns that are inevitably skewed reflections of reality, not direct reflections of it.

## 3. The Epistemic Challenges of Prediction

The failure of the ICEWS dataset to accurately measure Thailand's conflicts and networks gives rise to the question of why ICEWS researchers were apparently able to achieve relatively high accurate prediction rates for the onset of civil wars (Ward et al. 2013b), irregular leadership turnovers (Beger et al. 2014), and various types of conflicts (Ward 2016) based on the ICEWS dataset. This goes to the heart of the broader question of whether epistemic challenges are truly a hindrance to predicting human behavior.

To answer them will require us first to step back from Thailand in search of a wider theoretical framework for the discussion. I will draw on the four forecasting quadrants posited by Nassim Taleb (2004, 2007, and 2008; Taleb and Pilpel 2004), who argues that the forecasting potential of statistics is limited when research questions fall into a world of uncertainty with complex outcomes and unknown population distribution parameters in the tails. Then, returning to the forecasting endeavors of the ICEWS team, I will show that they achieve predictive success only by replacing a difficult research agenda that borders on the world of uncertainty into an easy

prediction task for which statistics work well. This comes at the cost of depriving their predictions of scientific value.

That this should be the case is congruent with Philip Tetlock's famous study of the accuracy of political forecasts made by experts, undergraduates, and simple and sophisticated statistical models. Tetlock (2005) showed that experts performed only slightly better in forecasting than did undergraduates or randomness, while the style of reasoning used by experts, which he dichotomizes into "hedgehogs" and "foxes," made a substantial difference. Hedgehogs are experts in one theory and tend to apply it universally for predictions on all issues. By contrast, foxes know many small things, they are more willing to adjust their position upon getting new information, and they use ad-hoc reasoning to make predictions. Tetlock (2005) finds that foxes significantly outperform hedgehogs. However, this does not mean that foxes were generally good forecasters. While the best foxes could predict about 20 percent of the variance, simple statistical models reached accuracy levels of 25 to 30 percent and an autoregressive distributed-lag model explained 47 percent (Tetlock 2005, 53).

Tetlock's study suggests that radical skepticism about forecasting is unwarranted. Open-minded forecasters performed systematically better than their ideologically constrained counterparts, and sophisticated statistical models produced a good record by inferring from patterns of the past to the near future. But many political events were generally unforeseen, amongst them seminal events such as 9/11 and, later, the Arab Spring.

More recently, many scholars have become convinced that the growing availability of large-scale datasets with unprecedented accuracy will lead to better predictions by experts armed with statistical models. Gary King (2009, 91-93) maintains that "our knowledge of practical solutions for problems of government and politics will begin to grow at an enormous rate – if we

are ready. […] Hundreds of years from now, social science today will look like it did when they first handed out telescopes to astronomers."

The key assumption behind the belief that big data will substantially improve predictions is the usual statistical assumption: that a random sample will be representative of a population's real parameter distribution. The larger the random sample gets, the faster it should approach the Gaussian normal distribution. Thus, the sample estimator should get closer to the real mean and real standard deviation when more and better data from the past are available. Statistical predictions indeed work well for a population with a stable Gaussian normal distribution. But our confidence in predictions should decline when the population is prone to an unexpected event that drastically changes its stationarity (Taleb 2004, 112-13). Taleb (2007, xvii–xviii) calls such an unexpected outlier a "black-swan event" if it has a strong substantial impact and if no past information could reasonably point to its occurrence.

Taleb does not argue that the world is generally unpredictable, but that the reliability of predictions made by statistical models depends on the probability characteristics of the area in which they are applied. By using a 2 x 2 matrix, he distinguishes between four ideal-type probability quadrants based on the population distribution and payoff structure, which are depicted by Table 1.

**<<< TABLE 1 >>>**

Whereas simple payoffs refer to binary "yes" and "no" distinctions (e.g., will a war occur?), complex payoffs consist of several continuous outcomes (e.g., how many casualties will occur in a war?) and different outcome impacts. A different outcome impact means that we do not know ex

ante whether the face value of an outcome will be the same as its impact value. For instance, a fall in stock returns of 2 percent might routinely not affect the broader behavior of investors by more than 2 percent, but in some cases it could cause panic and herding behavior that leads to a far greater fall in returns.

Distribution 1 refers to a scenario in which the probability generator of a distribution is observable, thus to a world of calculable risks rather than uncertainty. The population approximates Gaussian normal distribution. Statistical forecasts perform well for Distribution 1, particularly for the simple, binary outcomes in Quadrant 1. Larger and better data tend to make statistical predictions more accurate in Quadrants 1 and 2, as information from the past is (in these quadrants) meaningful for predicting the future.

Distribution 2 is a world with heavy or unknown tails in which the outcomes are visible but the probability generator is concealed. The heavy tails indicate that we are moving into a world of uncertainty where we can see the result (say, the outcome of a roll of the dice) but do not know what produced it (e.g., how many sides the dice have). Statistical predictions will encounter difficulties in Distribution 2 if the probability generator produces outcomes that differ from the past. Statistical predictions still tend to work, at least in the short run, for binary outcomes in Quadrant 3 of Distribution 2, because errors are constrained by the need to predict one of only two possible outcomes. Thus, the Law of Large Numbers is still at work, and extreme deviations do not carry a strong impact (Taleb and Tetlock 2013). Unfortunately, Quadrant 3 is often a function of categories deliberately created by human beings, including the observer-predictor. There are not many simple decisions in the real world of uncertainty (Taleb 2008).

In Quadrant 4 the probability distribution is concealed and payoffs are complex. Thus, outliers do matter for the sample distribution. The concealed probability generator creates new

causes of outcomes, and neither their emergence nor their impact could be predicted from past observations—regardless of data size or quality. The evolutionary process or financial markets appear to often fall into Quadrant 4. It was impossible from past observation to predict the emergence of financial innovations, such as mortgage-backed securities, and how they would interact with other elements of the financial system. An unexpected event in financial markets can have an extreme and lasting impact that renders the stationarity of an existing time series irrelevant. Thus, a decade-long time-series for the Western banking systems provided a quite accurate estimation of average returns of profits, but only until the financial meltdown of 2007-2008 blew up the historical trend (Blyth 2009, 452; 457-58).

Mark Blyth (2006, 497) argues that political science cannot generally be in the first or fourth quadrants. If it would be in Quadrant 1, scientific inquiry would be easy, but many of our theories are contestible (and are contested). By contrast, if political science were in Quadrant 4, all social outcomes would be unpredictable, which is also not the case. Political science tends to be regularly situated in between, because people create institutions that impose a certain degree of stability on an uncertain world. Paul Pierson (2000, 253) illustrates the stability and path dependency of human institutions by using Pólya's urn example. An urn contains two balls, red and green. A player draws a ball at random and returns the chosen ball plus an additional ball of the same color after every draw. The initial outcomes are unpredictable, but as every draw involves positive, reinforcing feedback, the outcome will eventually settle down on a stable equilibrium. Statistical forecasts and game-theoretic models are relatively successful in predicting outcomes in judicial or legislative politics because the probability generators and outcomes are human-made and often directly observable, rules are given, votes can be counted, and outliers do not tend to affect the underlying stationarity (Blyth 2006, 496).

The simplifying effect of institutions also explains the success of Nate Silver (2012) and other forecasters in predicting election results. The first-past-the-post Electoral College system of U. S. Presidential elections effectively turns the election into a two-candidate race with a few highly contested states, for which it is relatively easy to estimate reliable probabilities based on accurate opinion polls shortly before an election. The accuracy of election predictions declines, however, if a systematic polling error occurs or elections become more complex. Forecasters failed to predict the Brexit vote—a simple nationwide 'yes' or 'no' decision—due to systematic polling errors that underestimated the exit-vote. Similarly, all major forecast models failed to predict Trump's victory in the 2016 U. S. Presidential election due to polling errors. The 2015 British parliamentary election shows that predicting the election outcome in a first-past-the-post system based on single districts with multiple effective contenders proves to be difficult—and not just because opinion polls fell short of sufficiently capturing vote intentions (Booth 2015).[4] Furthermore, an unexpected event not covered by one's election model can drastically distort predictions of voting behavior, as illustrated by the 2004 Spanish parliamentary election. Although opinion polls predicted a victory for the conservative government shortly before the election, Islamist train bombings three days before the election, as well as the government's public handling of the attacks, led to a substantial change in public opinion and a surprise electoral victory of the Spanish social democrats. The Spanish case highlights the fact that human-made institutions are still situated in a world of uncertainty.

Statistical models are well equipped to prepare us for a future that behaves like the past, but they encounter problems when facing extreme events in the tails. As such events occur rarely (by definition), statistical models generally tend to have good forecasting capabilities, particularly

[4] Nate Silver's team also failed in their prediction for the UK election. None of their prediction intervals included the true number of seats for the four biggest parties (Lauderdale 2015).

if forecasts are transformed into binary-choice problems. But black swans, which are often the driving force of history, occur more often than suggested by a Gaussian normal distribution when there are concealed probability generators. More and better data from the past will not make unknown unknowns detectable in advance. To the contrary, the same technological advances that made the big-data revolution possible might also foster market and regulatory homogenization, thereby increasing susceptibility to black-swan events (Taleb 2007, 225-26).

Pasquale Cirillo and Taleb's statistical analysis of conflicts over common era history finds that violent conflicts have a fat-tailed Quadrant-3 or -4 distribution with "memoryless inter-arrival times, thus incompatible with the idea of a time trend" (Cirillo and Taleb 2016). Other studies show that the distribution of casualties in human conflicts has heavy tails, and often resembles a power law distribution, i. e. the probability of an event is inversely related to its impact, which falls into the category of Distribution 2 (Clauset et al. 2007; Scharpf et al. 2014; Friedman 2015).

Wars generally qualify for the complex payoff and fat tails distribution of Table 1: The death toll of wars could have multiple outcomes – potentially reaching billions in the extreme case of a nuclear war, which potentially have a varying impact, as a certain number or type of casualties could draw outside powers into the conflict. This would place the study of casualties of wars into Quadrant 4, where accurate predictions are likely to remain elusive.

## 4. An Analysis of the ICEWS Project

I will return now to the ICEWS project to explain its apparent success in predicting the future.

First, notice that Ward et al. (2013b) distinguish between 'onset of war' and 'no war', thus turning the complexity of war casualties into a binary research agenda that is situated in Quadrant 3. Binary statistical predictions in a Quadrant-3 world could potentially have some degree of

accuracy because the error rate is bounded. As the ICEWS provides large-scale event data on a daily basis, Ward and colleagues abandon the traditional annual format of time-series analysis to conduct forecasts on a monthly basis for the following six months. They incorporate low and high intensity ICEWS event variables in their model, measuring the behavior between the government and opposition.

**<<< TABLE 2 >>>**

Apparently this model performs well, as shown in Table 2. When the probability threshold given by the model is 50 percent or higher for the occurrence of a civil war, the model correctly predicts 199 of 286 civil wars (69.6 percent), while falsely predicting a civil war that did not occur in 33 cases. As the threshold is reduced, the number of correctly predicted civil wars increases, as does the number of false positives. For a threshold of 5 percent or higher, there are 424 civil war forecasts, of which about 61.6 percent turn out to be correct, capturing 91.3 percent of all civil wars. Similarly, the forecasts of irregular leadership changes made by Beger et al. (2014), based on ICEWS event data, also appear to be successful for the Ukraine and Thailand. A six-month out-of-sample predictions for 167 countries each month produce accurate prediction rates of up to 98 percent for ethnic religious violence, 96 percent for dyadic international crisis, 89 percent for insurgency, 84 percent for rebellion, and 68 percent for domestic political crisis (Ward 2016, 86). Based on such results (and echoing King 2009), Ward et al. (2013b, 486-87) conclude that, "given advancements in theory, data collection, statistics, and computational power, we might be at an important point to push the boundaries of predicting political phenomenon beyond what we

believed was possible only a few years ago. To preemptively declare defeat at the forecasting task seems foolish."

Let us pause for a moment to ask why prediction is so important to these social scientists. Ward et al. (2013b, 480) contend that making accurate predictions is a "gold standard" for model evaluation and scientific advancement, because "good explanations based on good social science theory should be able to generate accurate predictions, even if they will be probabilistic." Focusing on predictions rather than ex-post explanations is also supposed to solve an incentive problem facing researchers. Using ex-post analyses, researchers would tend to choose their preferred theory, "often focusing on the novel or unusual," to explain statistically significant variables. In contrast, a predictive social scientist has the incentive to integrate different parts of theories that actually work, as his explanation will be later judged on how well it was able to predict the future.

But such predictive heuristics would be a scientific improvement only in cases where reliable predictions are epistemically possible. Apparently, the ICEWS researchers assume either that we live in a world that mostly resembles Quadrant 1, or that statistical techniques can succeed in correctly predicting events that often fall into Quadrant 4.

## 5. The Non-Universality of Game-Theoretic Abstractions

Regarding the theoretical underpinnings of their predictions, the models rely on standard game theory. Game-theoretical models rest on the assumption that it is an easy epistemologically task for rational self-interested agents to interpret their structurally given interest in any environment. But while structurally given interest is certainly an important factor, it does not come with an instruction sheet (Blyth 2003). It has to be mediated through media sources, and political ideas affect the interpretation of new information. For instance, without access to critical media sources,

voters will not know about the corruption charges against a popular politician; even if an impoverished voter receives reports on growing economic inequality, he would not develop discontent when holding a political belief that does not equate economic inequality with injustice (e. g., Jäger 2012).

The possibility that individual behavior indeed departs from the postulated behavior of structural models highlights the relevance of Weberian *Verstehen* for modern empirical social science research – the understanding of specific agents' intentions before one can engage in theoretical abstractions.

The problems of applying a universal game-theoretic model to a particular case without Weberian *Verstehen* can be illustrating by examining the game-theoretic model that was applied for Thailand. Metternich et al. (2013) argue that if a group succeeds in toppling a government, it will enact policies that are public goods for groups with a similar ideology. These ideologically similar groups therefore have no incentive to contribute to anti-government efforts that might topple the old government in the first place. This free rider incentive is supposed to ensure that conflicts are more prevalent when the anti-government network is ideologically incoherent and fragmented. Thus, contrary to what one might think, conflict is the result of a disunified opposition, not a unified one.

This model requires that it is epistemically easy for groups to know whether another group is able to overthrow the government. If there is uncertainty about the success of this endeavor, an ideologically similar group would not have the incentive to free ride, because it might, for all it knows, be necessary for it to join the anti-government effort if the government is to be toppled and the desired public goods provided.

None of the model's assumptions would seem useful in capturing actors' motivation in the Thai context. First, in explaining participation in mass protests, the free-rider theory runs into the same problem as when it is used to explain mass voting. The probability that personal participation will make any difference is virtually zero while the personal costs are clearly higher, leading to the rational expectation that participation should be rare. Second, programmatic orientation is not the only basis for party-citizen linkages. Especially in non-postindustrial democracies, selective material benefits are often motivating factors for voters and politicians (Kitschelt 2000). Apart from the core groups in the red-yellow divide, political parties in Thailand neither develop mass memberships nor programmatic agendas. These parties are often family-run election machines, frequently switching parties or government coalitions to increase the material benefits of their leaders (Croissant and Völkel 2012; Ufen 2012). For instance, the "Friends of Newin Group" defected from the pro-Thaksin party and its governing coalition in December 2008 to form the Bhumjaithai (Thai Pride) Party, part of a new anti-Thaksin Democrat-led government coalition. The leaders of political groups such as this one are regularly rewarded with patronage and higher administrative positions if they participate in a successful turnover.

A standard response to the criticism that theoretical explanations are unrealistic is to evoke Milton Friedman's postulate that theories can be unrealistic as long as their predictions prove to be correct (Friedman 1953). But this essentially turns theory-driven forecast models into merely data-driven predictions, because they cannot adequately explain *why* a certain variable is associated with a certain outcome.

In fact, previous qualitative and quantitative studies on Thailand's social movement mobilization falsify the game-theoretic model of a fragmented anti-government network conjured up by Metternich et al (2013). These studies show that conflicts were caused by factors unidentified

by the model, and that anti-government groups did indeed have to unite to effectively challenge the government through mass protests. Duncan McCargo (2005 and 2008), for example, shows that Thailand's political conflicts emerged out of power struggles between two competing elite networks. Prime Minister Thaksin employed state policies to break the power of network monarchy, inadvertently causing the escalation of the conflict in the deep South of the country and the emergence of a broad political movement against him. The pro- and anti-Thaksin movements in the color-coded conflict used mass rallies to challenge the government when they had the appropriate mobilization resources and when political events could unify their movement for a common purpose by means of political discourse and symbols (Pye and Schaffar 2008; Kengkij and Hewison 2009). Subsequently, the main political parties increasingly invested in their organizational capacity to increase the mobilization potential of their networks (Naruemon and McCargo 2011; Sinpeng and Kuhonta 2012; Sinpeng 2014).[5] Thus, the model in question appears to be irrelevant to recent collective action in Thailand, and may also be unrealistic if applied to other political environments.

## 6. Turning Civil-War Predictions into a Quadrant-1 Task

In light of all of this, it is remarkable that Ward et al. (2013b) apparently succeed in predicting up to over 90 percent of civil wars correctly. However, a closer look at their ICEWS event variables reveals that the methodology used by the ICEWS researchers transformed the research question from an epistemically difficult task of making predictions in Quadrant 3 into a much easier task with many characteristics of Quadrant-1 predictions.

[5] These studies are ignored by Metternich et al. (2013) in its discussion on Thailand's conflicts.

Ward et al. (2013b) use the Uppsala Conflict Data Program (UCDP) definition of the onset of a civil war as dependent variable. The UCDP codebook states that "a conflict is in the UCDP defined as starting when there is an open challenge by a state against another state, or an opposition group against a state, to resort to violence to reach its goal." If such a "stated goal of incompatibility" is absent or not detectable due to poor media coverage, the other definitions for the starting date are the use of military force, the occurrence of the first battle-related casualties, or the threshold of 25 battle-related deaths.[6] In order to explain the onset of UCDP-defined civil wars, Ward et al. (2013b) rely on the two ICEWS event variables labeled "behavioral." These variables measure whether high-intensity events, such as protests, fights, or killings, and low-intensity events, such as demands or threats, are occurring between pro- and anti-government groups in news reports (Ward et al. 2013b, 483).

This approach drastically simplifies the epistemic difficulties of forecasting by exploiting differences in artificial definitions. "Threats," "Fights," "Killings," and other violent activities are not independent of the constructed measurement of civil war. To the contrary, the UCDP measurement of civil war appears to be just a slightly more intense measurement of conflict than these two ICEWS variables, and thus a variation of the same thing.[7] As most civil wars are preceded by at least some form of tension, violence, or protest, it is not surprising that the ICEWS variables, which measure news reports of tensions, violence, and protests, are strongly associated with the occurrence of civil wars. This is represented by the ability of Ward et al. (2013b)'s model

[6] The UCDP definitions can be found at http://www.pcr.uu.se/research/ucdp/definitions.

[7] Moreover, Ward et al. (2013b)'s dataset is based on monthly data for forecasts over a six-month period. This reduces the difficulty of making predictions even further, as the escalation phase of a conflict usually lasts longer than a month before it is coded as a civil war. It is thus imaginable that the UCDP conflict count is already close to the 25-death threshold for a civil war in a given month, allowing for a relative easy prediction based on current trends as to whether the threshold will be surpassed in the next six months.

specification to capture 261 of 286 civil wars at the lowest probability threshold, suggesting that journalists reported some form of tension before a conflict was coded as a civil war. However, many conflicts reported in the news do not turn into a civil war, as illustrated by the fact that 163 of the model's civil-war predictions were false positives. Furthermore, the statistical model is blind to "black swans" that turn into a civil war without a tension phase, and cases for which the preceding tension phase is overlooked by news reports. This is to say that the 25 false negatives were probably far more difficult to predict but, for that very reason, far more relevant to those who seek to predict conflicts that are not already on everyone's radar screens. To take up a meteorological example, if one were to define tornadoes as including all periods of windy weather, the "tornado" prediction rate would increase, but the tornadoes from which we seek protection would be just as hard to predict as before. Important predictions tend to be difficult ones almost by definition; "predicting the future" becomes much easier if the predictions are banal.

A study in the *Journal of Peace Research* by Thomas Chadefaux (2014) is another example of research that achieves predictive success, in this case for interstate wars, by relying on news reports. Chadefaux based his research on a simple search count of country names and various keywords in the Google News Archive, such as "tension," "conflict," and "combat," yielding a weekly database for the period 1902-2001. He found that conflict-related news skyrocketed immediately before the onset of an interstate war, and showed that news reports anticipated the coded occurrence of wars with an accuracy rate of 85 percent within the next three months. The success rate of Chadefaux (2014, 15)'s "imperfect, ad hoc, and simple" news count without a network dimension illustrates why it is possible that the ICEWS database failed to generate accurate networks for Thailand in Metternich et al. (2013), while achieving relatively high accuracy in Ward et al. (2013b), Beger et al. (2014), and Ward (2016). In essence, the ICEWS dataset is able to

capture news reports announcing that tensions or skirmishes have occurred, which in turn can "predict" conflicts in the short-run.

To use Taleb's typology, the predictions by the ICEWS researchers are not difficult because they tend to fall into Quadrant 1. Journalism-based variables tend to transform the underlying fat tail distribution of war casualties into a Gaussian normal distribution that is truncated at the defined conflict threshold. Based on the current rate of reported tensions and violence, a statistical model projects this trend into the future, providing accurate probabilities for the possibility of reaching the conflict threshold in a given period. As a consequence, the statistical models reach very high levels of prediction accuracy that are typical for phenomena that mostly exhibit Quadrant 1 dynamics. It is also doubtful, then, that the ICEWS project would, as advertised, outperform country experts in making accurate predictions. Country experts and politically interested journalists should be aware of the possibility of a civil war or an irregular government turnover by following news reports, which is what ICEWS did.

This can be illustrated by the correctly predicted leadership changes in the Ukraine (probability of 28 percent) and Thailand (probability of 6 percent) for the period April-September 2014 in Beger et al. (2014). In the Ukraine, President Yanukovych was ousted on 22 February 2014 and a presidential snap election was announced shortly thereafter for 25 May 2014; until then, there was an acting president. So it was not difficult to anticipate in March that there would be another leadership turnover before September. In Thailand, political instability prevailed after October 2013, when months-long anti-government protests erupted, demanding the resignation of Thaksin's sister, Prime Minister Yingluck Shinawatra. The opposition movement boycotted the snap elections of 4 February 2014. By the end of March, the Constitutional Court had annulled the election, Yingluck was facing a trial in May over a rice-distribution scheme, and anti-government

protests still loomed. Given Thailand's history of frequent coups, and the fact the military and the judiciary had previously overthrown pro-Thaksin governments, an irregular removal of Yingluck within the next six months was not unlikely. In fact, country experts and journalists were overtly discussing this possibility, as summarized by André Vltchek's article "Thailand in danger: Watch out for yet another coup" (Vltcheck 2014), or by the *New York Times*, which carried a headline on 31 March 2014 stating that "In Thailand, Some Foresee a Coup by Legal Means" (Fuller 2014).

The scientific value of the ICEWS prediction project would seem to be comparable to a statistical model that achieves a nearly perfect prediction record for a presidential election by relying on a set of variables that measure the respondents' partisan identity and candidate ratings. But attributes such as being a Democrat or liking Trump very much are "variants of the very thing expressed by their votes on Election Day" (Friedman 2012, 309). The scientifically fruitful and demanding endeavor would be to explain why people hold such attitudes, or whether other, less obvious variables could predict their behavior. We do not learn much about the origins of conflicts from the insight that it is likely that a greater level of violent conflict will occur if newspapers have recently reported on lower levels of violent conflict. The reported phenomena themselves, e.g., why mass protests turn violent or require a government's resignation, are the crucial outcomes that require scientific inquiry. That civil wars occur is *explained* neither by the fact that they happen more often if journalists report on political violence shortly beforehand, nor by labeling news-coded variables as "behavioral" or "network" variables supported by a universalistic game-theoretic model.

## 7. Conclusion

While Ward et al. (2013b) claim that predictions create the right incentives for researchers, their own example rather confirms Taleb and Tetlock (2013)'s remark that there is a tendency to replace epistemically difficult predictions with easy ones. Taleb and Tetlock were referring not to the placement of weaker measurements of the dependent variable on the right-hand side of the regression equation, but to the practice of turning a Quadrant-4 question into a Quadrant-3 (or even -1) question by creating binary categories for predictions. Such "well-intentioned efforts to improve performance in binary prediction tasks can have the unintended consequence of rendering us oblivious to catastrophic variable exposure. […] Framing a 'variable' question as a binary question is dangerous because it masks exponentially escalating tail risks, the risks of a confrontation claiming not just 10 lives [but] 1000 or 1 million." This is a particular problem for the ICEWS project, which is intended to inform U. S. policy makers about potential conflicts. It is well-designed to estimate from recent news articles the probability that a country will be coded as a conflict case, thus whether scattered conflicts increase to a defined conflict threshold in the next six months. But it is blind to the severity of conflicts that go beyond the threshold.

If there is a larger lesson here, however, it seems to me that it regards the equation of science with prediction. Science is the search for truth using empirical tests and reasoned skepticism about claims. We have little reason to think that the process of scientific inquiry about human behavior will yield predictive formulae, but this is because of the unique nature of human (cognitively directed) behavior, not because of the absence of enough data.

**References**


Beger, Andreas, Cassy L. Dorff, and Michael D. Ward. 2014. "Ensemble Forecasting of Irregular Leadership Change." *Research & Politics* 1(3) (November): 1-7.

Blyth, Mark. 2003. "Structures Do Not Come with an Instruction Sheet: Interests, Ideas, and Progress in Political Science." *Perspectives on Politics* 1(4) (December): 695-706.

Blyth, Mark. 2006. "Great Punctuations: Prediction, Randomness, and the Evolution of Comparative Political Science." *American Political Science Review* 100(4) (November): 493-98.

Blyth, Mark. 2009. "Coping With the Black Swan: The Unsettling World of Nassim Taleb." *Critical Review* 21(4): 447-66.

Booth, Robert. 2015. "Why Did the Election Pollsters Get It So Wrong?" *The Guardian*, 14 May.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael D. Ward. 2015. *ICEWS Coded Event Data*. http://dx.doi.org/10.7910/DVN/28075. Harvard Dataverse, V5.

Brandt, Patrick T., John R. Freeman, and Philip A. Schrodt. 2014. "Evaluating Forecasts of Political Conflict Dynamics." *International Journal of Forecasting* 30(4) (October-December): 944-62.

Chadefaux, Thomas. 2014. "Early Warning Signals for War in the News." *Journal of Peace Research* 51(1) (January): 5-18.

Cirillo, Pasquale, and Nassim Nicholas Taleb. 2016. "On the Statistical Properties and Tail Risk of Violent Conflicts." *Physica A, Statistical Mechanics and Its Applications* 452 (June): 29-45.

Clauset, Aaron, Maxwell Young, and Kristian Skrede Gleditsch. 2007. "On the Frequency of Severe Terrorist Events." *Journal of Conflict Resolution* 51(1) (February): 58-87.

Croissant, Aurel, and Philip Völkel. 2012. "Party System Types and Party System Institutionalization: Comparing New Democracies in East and Southeast Asia." *Party Politics* 18(2) (March): 235-65.

Friedman, Jeffrey. 2012. "System Effects and the Problem of Prediction." *Critical Review* 24(3), 291-312.

Friedman, Jeffrey A. 2015. "Using power laws to estimate conflict size." *Journal of Conflict Resolution* 59(7) (October): 1216-41.

Friedman, Milton. 1953. "The Methodology of Positive Economics." In idem, *Essays in Positive Economics*. Chicago: University of Chicago Press.

Fuller, Thomas. 2014. "In Thailand, Some Foresee a Coup by Legal Means." *New York Times*, 31 March.

Goldsmith, Benjamin E., Charles R. Butcher, Dimitri Semenovich, and Arcot Sowmya. 2013. "Forecasting the Onset of Genocide and Politicide: Annual Out-Of-Sample Forecasts on a Global Dataset, 1988–2003." *Journal of Peace Research* 50(4) (July): 437-52.

Hegre, Håvard, Joakim Karlsen, Håvard M. Nygård, Håvard Strand, and Henrik Urdal. 2013. "Predicting Armed Conflict, 2011–2050." *International Studies Quarterly* 57(2) (June): 250–70.

Jäger, Kai. 2012. “Why Did Thailand’s Middle Class Turn against a Democratically Elected Government? The Information-Gap Hypothesis.” *Democratization* 19(6) (December): 1138-65.
Jäger, Kai. 2016. “The Limits of Studying Networks via Event Data: Evidence from the ICEWS dataset.” Working Paper.
King, Gary. 2009. “The Changing Evidence Base of Social Science Research.” In *The Future of Political Science*, ed. Gary King, Kay Schlozman, and Norman Lie. London: Routledge.
Kitschelt, Herbert. 2000. “Linkages between Citizens and Politicians in Democratic Polities.” *Comparative Political Studies* 33(6-7) (September): 845-79.
Kengkij, Kitirianglarp, and Kevin Hewison. 2009. “Social Movements and Political Opposition in Contemporary Thailand.” *The Pacific Review* 22(4): 451-77.
Lauderdale, Ben. 2015. “What We Got Wrong In Our 2015 U.K. General Election Model.” http://fivethirtyeight.com/datalab/what-we-got-wrong-in-our-2015-uk-general-election-model.
McCargo, Duncan. 2005. “Network Monarchy and Legitimacy Crises in Thailand.” *The Pacific Review* 18(4): 499-519.
McCargo, Duncan. 2008. *Tearing Apart the Land: Islam and Legitimacy in Southern Thailand*. Ithaca: Cornell University Press.
Metternich, Nils W., Cassy Dorff, Max Gallop, Simon Weschle, and Michael D. Ward. 2013. “Antigovernment Networks in Civil Conflicts: How Network Structures Affect Conflictual behavior.” *American Journal of Political Science* 57(4) (October): 892-911.
Naruemon, Thabchumpon, and Duncan McCargo. 2011. “Urbanized Villagers in the 2010 Thai Redshirt Protests.” *Asian Survey* 51(6) (November/December): 993-1018.
Nidhi, Eoseewong. 2012. “The Culture of the Army. Matichon Weekly, 28 May 2010.” In *Bangkok, May 2010: Perspectives on a Divided Thailand*, eds. Michael J. Montesano, Pavin Chachavalpongpun, and Aekapol Chongvilaivan. Singapore: Institute of Southeast Asian Studies.
Pierson, Paul. 2000. “Increasing Returns, Path Dependence, and the Study of Politics.” *American Political Science Review* 94(2) (June): 251-67.
Pye, Oliver, and Wolfram Schaffar. 2008. “The 2006 Anti-Thaksin Movement in Thailand: An Analysis.” *Journal of Contemporary Asia* 38(1): 38-61.
Scharpf, Adam, Gerald Schneider, Anna Nöh, and Aaron Clauset. 2014. “Forecasting the Risk of Extreme Massacres in Syria.” *European Review of International Studies* 1(2) (August): 50-68.
Schrodt, Philip A, and David Van Brackle 2013. “Automated Coding of Political Event Data.” In *Handbook of Computational Approaches to Counterterrorism*, ed. V.S. Subrahmanian. New York: Springer.
Silver, Nate. 2012. *The Signal and the Noise*. New York: Penguin.
Sinpeng, Aim. 2014. “Party-Social Movement Coalition in Thailand’s Political Conflict (2005–2011).” In *Contemporary Socio-Cultural and Political Perspectives in Thailand*, ed. Pranee Liamputtong. New York: Springer.
Sinpeng, Aim, and Erik Martinez Kuhonta. 2012. “From the Street to the Ballot Box: The July 2011 Elections and the Rise of Social Movements in Thailand.” *Contemporary Southeast Asia* 34(3) (December): 389-415.
Taleb, Nassim Nicholas. 2004. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. New York: Random House.

Taleb, Nassim Nicholas. 2007. *The Black Swan: The Impact of the Highly Improbable*. London: Penguin.
Taleb, Nassim Nicholas. 2008. "The Fourth Quadrant: A Map of the Limits of Statistics." Unpublished Manuscript. http://www.edge.org/3rd_culture/taleb08/taleb08_index.html
Taleb, Nassim Nicholas, and Avital Pilpel. 2004. "On the Unfortunate Problem of the Nonobservability of the Probability Distribution." Unpublished Manuscript. http://www.fooledbyrandomness.com/knowledge.pdf
Taleb, Nassim Nicholas, and Philip E. Tetlock. 2013. "On the Difference between Binary Prediction and True Exposure with Implications for Forecasting Tournaments and Decision Making Research." Unpublished Manuscript. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2284964
Tetlock, Philip E. 2005. *Expert Political Judgment*. Princeton: Princeton University Press.
Ufen, Andreas. 2012. "Party Systems, Critical Junctures, and Cleavages in Southeast Asia." *Asian Survey* 52(3) (May/June): 441-64.
Ulfelder, Jay. 2012. "Forecasting Onset of Mass Killings." Paper presented at the Annual Northeast Political Methodology Meeting at New York University
Vltchek, André. 2014. "Thailand in Danger: Watch Out for Yet Another Coup." *Russia Today*, 28 February.
Ward, Michael D. 2016. "Can We Predict Politics? Toward What End?" *Journal of Global Security Studies* 1(1) (February): 80-91.
Ward, Michael D., Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013a. "Comparing GDELT and ICEWS Event Data." Working Paper. http://mdwardlab.com/sites/default/files/GDELTICEWS_0.pdf
Ward, Michael D., Nils W. Metternich, Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, and Simon Weschle. 2013b. "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction." *International Studies Review* 15(4) (December): 473-90.
Ward, Michael D., and Nils W. Metternich. 2012. "Predicting the Future Is Easier Than It Looks." *Foreign Policy*, 12 November.

**Table 1: The Four Probability Quadrants and the Forecasting Power of Statistics**

| | simple payoffs<br>(binary predictions) | complex payoffs<br>(variable predictions) |
|---|---|---|
| Distribution 1<br>(thin tailed) | *Quadrant 1*<br>Statistical forecasts do very well. | *Quadrant 2*<br>Statistical forecasts do well. |
| Distribution 2<br>(heavy and/or unknown tails) | *Quadrant 3*<br>Statistical forecasts are often reliable. | *Quadrant 4*,<br>The limit of statistics: statistical forecasts do not tend to work. |

*Source:* Taleb 2008.

**Table 2: Performance of the Ward et al. (2013b) Model in Predicting the Onset of Civil Wars**

| Prob. Threshold | Civil War Forecasts | Correctly Predicted | False Positives |
|---|---|---|---|
| 0.5 | 232 | 199/286 | 33/1781 |
| 0.3 | 293 | 232/286 | 61/1781 |
| 0.1 | 362 | 245/286 | 117/1781 |
| 0.05 | 424 | 261/286 | 163/1781 |

*Source*: Ward et al. 2013b, 485.